

Comparative analysis of rigidity across protein families

S A Wells¹, J E Jimenez-Roldan^{1,2} and R A Römer¹

¹ Department of Physics and Centre for Scientific Computing, University of Warwick, Coventry, CV4 7AL, United Kingdom

² Department of Systems Biology, University of Warwick, Coventry, CV4 7AL, United Kingdom

E-mail: r.roemer@warwick.ac.uk

Abstract. *Revision* : 1.3, compiled 3 June 2009

We present a comparative study in which “pebble game” rigidity analysis is applied to multiple protein crystal structures, for each of six different protein families. We find that the mainchain rigidity of a protein structure at a given hydrogen-bond energy cutoff is quite sensitive to small structural variations, and conclude that the hydrogen bond constraints in rigidity analysis should be chosen so as to form and test specific hypotheses about the rigidity of a particular protein. Our comparative approach highlights two different characteristic patterns (“sudden” or “gradual”) for protein rigidity loss as constraints are removed, in line with recent results on the rigidity transitions of glassy networks.

PACS numbers: 87.14.E-, 87.15.La

Submitted to: *Phys. Biol.*

1. Introduction

It is a common goal in biophysics to represent the flexibility of a protein and study its large-scale motion without incurring the full computational cost of molecular dynamics simulations. One popular family of approaches is based on normal-mode analysis applied to a full or simplified representation of the protein structure [1–10], with the aim of representing large-scale conformation change in terms of a reduced set of low-frequency motions [11]. Another approach is to divide up the protein structure into relatively rigid sections or domains, connected together by flexible regions or “hinges”. This can be done using a variety of structure-based approaches [12–17].

In this paper we concern ourselves with the “pebble game” [18], an integer algorithm for rigidity analysis. By matching degrees of freedom against constraints, it can rapidly divide a network into rigid regions and floppy “hinges” with excess degrees of freedom. The program FIRST implements this algorithm for protein crystal structures [19]. The rigid units in a protein structure may be as small as individual methyl groups or large enough to include entire protein domains containing multiple secondary-structure units. The division of a structure into rigid units is referred to as a Rigid Cluster Decomposition (RCD).

Rigidity analysis has been used to study phenomena such as virus capsid assembly [20] and protein folding [21, 22]. The coarse-graining provided by a RCD also forms the basis of simulation methods aiming to explore the large-amplitude flexible motion of proteins: the ROCK algorithm [23] and more recently the FRODA geometric simulation algorithm [24], which has been applied in various studies of protein flexibility [25–29], and the rigidity-enhanced elastic network model [30].

The results of rigidity analysis on proteins depend upon the set of constraints that are included, with the user setting an energy “cutoff” which determines the set of hydrogen bonds to include in the analysis, (see section 2). However, previous studies using FIRST have used widely differing, sometimes contradictory, cutoff values and methods of constraint selection — we give a brief review of the situation in Appendix A. This methodological issue not only makes it more difficult for scientists to adopt pebble-game rigidity analysis as a method, but also raises issues in the interpretation of results. There is at present no clear guidance on the “correct” choice of cutoff value; nor is it clear how comparable are the results of rigidity analysis using a given cutoff value on slightly different protein structures.

Hence the primary motivation for our study is to fill this gap by explicitly comparing the results of rigidity analysis on groups of very similar crystal structures. We concentrate particularly on eukaryotic cytochrome C while also considering five other proteins (hemoglobin, myoglobin, α -lactalbumin, trypsin and HIV-1 protease). For each protein structure we observe the pattern of rigidity loss during the progressive removal of hydrogen bonds, or “rigidity dilution” [21, 22]. We define *mainchain rigidity* as a measure of the rigidity of the protein backbone in order to describe the rigidity loss during dilution. On the basis of this study we comment on the selection of cutoff values and the interpretation of rigidity analyses.

The second motivation for our study is to observe the pattern of rigidity loss during dilution. Previous studies on protein folding [21] have drawn comparisons between the folding transition of proteins and the rigidity transition of glassy networks. A recent study [31] found that the rigidity transition in glasses could display either first-order or second-order behaviour depending on the character of the constraint network. In the first case, a small change in the constraints causes a sudden transition

from an entirely floppy state to one in which the entire system becomes rigid. In the second, rigidity develops in a percolating rigid cluster which initially involves only a small proportion of the network and then gradually increases in size as more constraints are introduced. Our data on rigidity dilution shows that both types of transition are possible in proteins, with four of our proteins typically displaying “gradual” rigidity change and two (trypsin and HIV-1 protease) displaying “sudden” rigidity change.

2. Materials and Methods

2.1. Protein selection

We have chosen sets of proteins from the protein data bank (PDB) [32] to obtain similar crystal structures for our comparison, as summarised in Table 1. We sought particularly (i) examples of the same protein from different organisms, e.g. cytochrome C proteins from multiple different eukaryotic mitochondria, and (ii) protein structures obtained under different conditions of crystallisation, e.g. in complex with different ligands, proteins or substrates. In the present study we will only investigate non-membrane proteins because the default treatment of hydrogen bonds and hydrophobic tethers in FIRST is based on the assumption that the protein exists in a polar solvent (cytoplasm) rather than being within a hydrophobic or amphiphilic environment as for membrane-bound proteins. Proteins in a membrane environment can still be handled but this requires hand-editing of the constraint network. Rigidity analysis is best carried out on crystal structures with high resolution, so that we can have confidence in the accuracy of the atomic positions when constructing the hydrogen-bond geometries. We therefore concentrated on X-ray crystal structures with resolutions of better than 2.5Å.

From each PDB crystal structure we extracted a single protein chain, eliminating all crystal water molecules, but retaining important hetero groups such as the porphyrin/heme units of cytochrome C and hemoglobin. The PYMOL visualisation software [33] proved very useful for this purpose. We add the hydrogens that are absent from X-ray crystal structures, using the REDUCE software [34] which also performs necessary flipping of side chains. After the addition of hydrogens we renumbered the atoms using PYMOL again to produce files usable as input to FIRST [19,21]. In the case of HIV protease we analysed the homodimer unit, as in [19].

2.2. Rigidity analysis and dilution

The energy of each potential hydrogen bond in the processed structure is calculated in FIRST using the Mayo potential [35]; the distance-dependent part of this potential is shown in Figure 1. For the dilution, FIRST performs an initial rigidity analysis including all bonds with energies of 0 kcal/mol or lower; bonds are then removed in order of strength, gradually reducing, or “diluting”, the rigidity of the structure.

An example of this rigidity dilution for a given protein is shown in Figure 2a for the 1HRC horse cytochrome C structure. The horizontal axis represents the protein’s linear primary structure. Flexible areas of the polypeptide sequence are shown as horizontal thin black lines while areas lying within a rigid cluster are shown as thicker coloured blocks. Colour is used to differentiate which residues belong to which rigid cluster. The three-dimensional protein fold makes it possible for residues that are widely separated along the backbone to be spatially adjacent and form a

Table 1. List of all the proteins, their organism of origin, PDB codes as well as the Figures in which they appear

Protein	Organism	PDB ID	Figure	Comments
Cytochrome C	Horse	1HRC	6	uncomplexed complexed with antibody E8 complexed with peroxidase at low ionic strength
		1WEJ		
		1U75		
		1CRC		
Cytochrome C	Tuna	5CYT	6	ferricytochrome 2FE:1ZN mixed-metal porphyrins 2ZN:1FE mixed-metal porphyrins Cobalt(III)-substituted
		1I54		
		1I55		
		1LFM		
Cytochrome C	Rice	1CCR	7a	
	Bonito	1CYC		
	Bacteria	1A7V		
	Tuna	1I55		
	Yeast	1YCC		
		2YCC		
Myoglobin	Horse	1DWR	7b	
	Whale	1HJT		
	Turtle	1LHS		
α -lactalbumin	Baboon	1ALC	7c	
	Human	1HML		
	Goat	1HFY		
	Human	1A4V		
	Guinea pig	1HFX		
	Cattle	1F6R		
Hemoglobin (α chain)	Human	1A3N	7d	deoxy oxy deoxy carbonmonoxy
		2DN1		
		2DN2		
		2DN3		
	Goose	1A4F		
	Rice	1D8U		
	Bacteria	1DLW		
	Alga	1DLY		
	Cattle	1G09		
	Worm	1KR7		
	Clam	1MOH		
HIV-1 Protease	Virus	1HTG	7e	homodimers with inhibitors bound
		4HVP		
		7HVP		
		8HVP		
		9HVP		
Trypsin	Salmon	1A0J	7f	
	Cattle	1AQ7		
		1AUJ		
	Pig	1AVW		
	Pig	1AVX		
	Cattle	1AZ8		
	Rat	1BRA		
		1BRB		
		1BRC		
	Cattle	1BTH		
	Salmon	1BZX		
	Human	1H4W		
		1HPT		
	Cattle	1K1I		
		1K1J		
		1K1M		
		1K1N		
		1K1O		
		1K1P		
	Pig	1LDT		
	Human	1TRN		
		2RA3		
	Rat	3TGI		

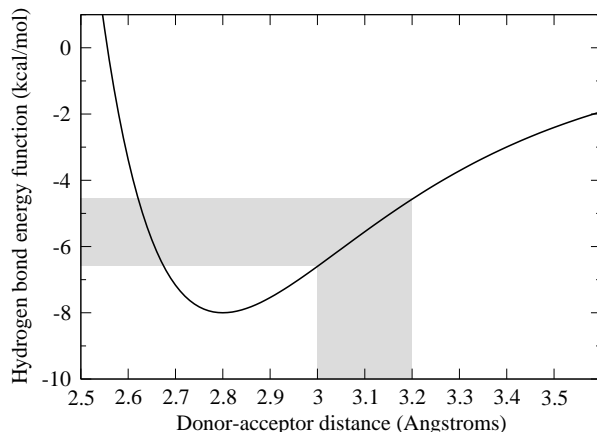


Figure 1. Dependence of hydrogen bond energy E in FIRST on the donor-acceptor distance. The shaded region indicates how an distance variation of $\pm 0.1 \text{ \AA}$ can lead to a variation in the bond energy of more than 1 kcal/mol.

single rigid cluster. The vertical axis on the dilution plot represents the dilution of constraints by progressively lowering the cutoff energy for inclusion of hydrogen bonds in the constraint network. Each time the rigid cluster analysis of the mainchain α -carbon atoms (C_α) changes as a result of the dilution, a new line is drawn on the plot, labelled with the energy cutoff and with the network mean coordination for the protein at that stage. We should stress that the RCD is always performed over the entire protein structure (mainchain and sidechain atoms) and a dilution is performed for every hydrogen bond removed from the set of constraints, typically several hundred bonds for a small globular protein. The dilution plot is then a summary concentrating on the rigid-cluster membership of the C_α atoms defining the protein backbone.

2.3. Mainchain rigidity loss during dilution

Dilution plots of very similar protein structures can be compared directly as shown in Figure 4. This form of comparison, however, becomes unwieldy when comparing large numbers of structures, and can obscure differences in the hydrogen-bond energy scale. For glassy networks [31] the overall degree of rigidity of the structure was measured by the number of atoms in the largest spanning rigid cluster in a network with periodic boundary conditions. Since the protein is not a periodic structure, we measure its overall rigidity by considering how many of its residues are included in large rigid clusters.

In Figure 3(a) we show the number n_N of C_α contained within the larger N rigid clusters of the horse cytochrome C structure 1HRC, for which the total number of C_α atoms equals $\mathcal{N}_{C_\alpha} = 105$. It is clear that only the first few rigid clusters (numbered 1–5) contain more than one C_α while higher-numbered clusters do not contain more than one C_α and do not represent two or more residues forming a single rigid unit. In Figure 3(b) we show the fraction f_N of C_α contained in the first N cluster, defined as

$$f_N(E) = \frac{1}{\mathcal{N}_{C_\alpha}} \sum_{i=1}^N n_i(E) \quad (1)$$

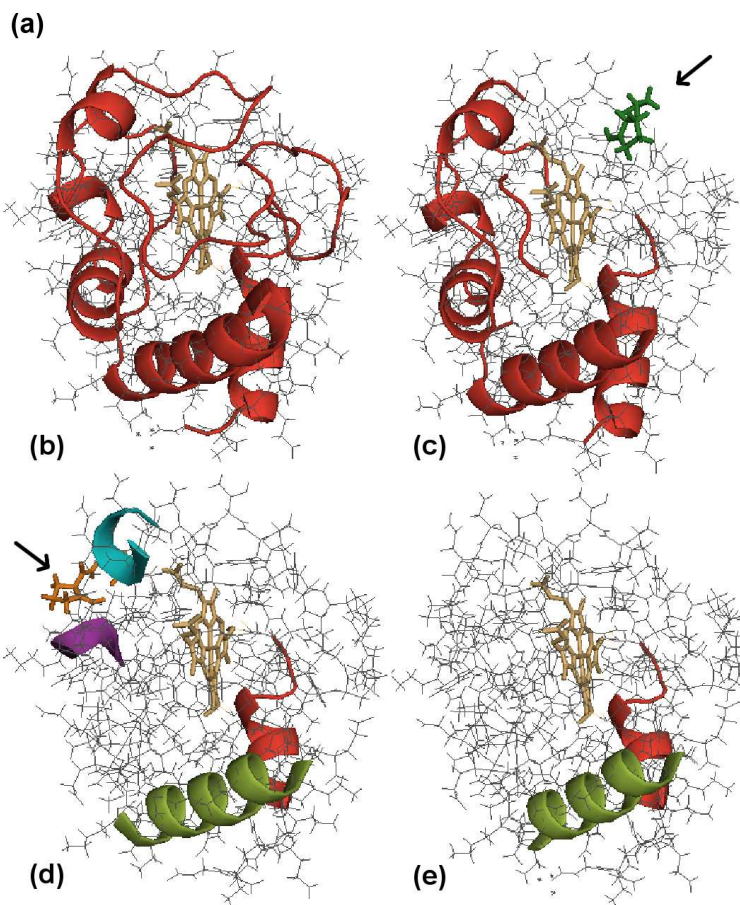
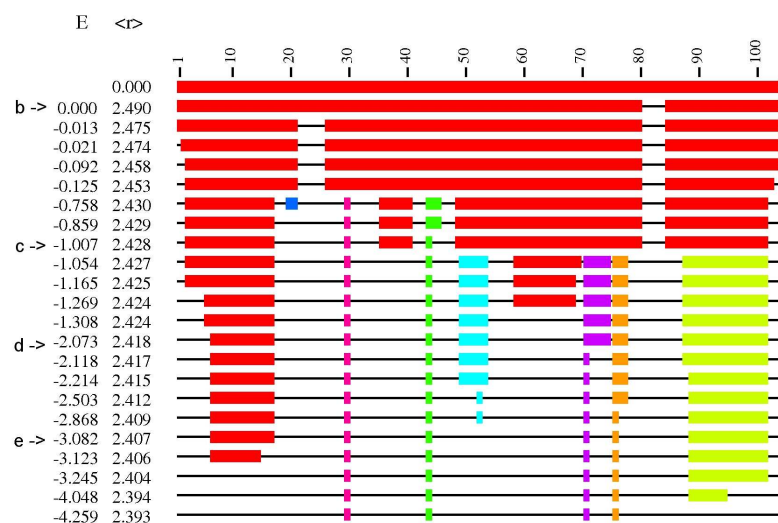


Figure 2. (a) Dilution plot for horse cytochrome C from the 1HRC structure. Flexible regions of the polypeptide chain appear as black thin lines, whereas rigid portions appear as coloured along the protein chain with C_α labelled from 1 to 105. The second column on the left indicates the mean number $\langle r \rangle$ of bonded neighbours per atom as the energy cutoff E changes. When E decreases (left-most column), rigid clusters break up and more of the chain becomes flexible. Colour coding shows which atoms belong to which rigid cluster. (b,c,d and e) Rigidity distribution for horse Cytochrome C from the 1HRC structure in 3D. These figures represent in grey the flexible regions and in colour the largest rigid regions for the native state at energy cutoffs (b) $E = 0.000$, (c) $E = 1.007$, (d) $E = 2.073$ and (e) $E = 3.082$, respectively. For each figure, the colour coding correlates with the colour coding given in (a). The arrows in (c) and (d) indicate two smaller rigid clusters shown in “stick” representation for clarity. The heme group is shown in “stick” representation (yellow).

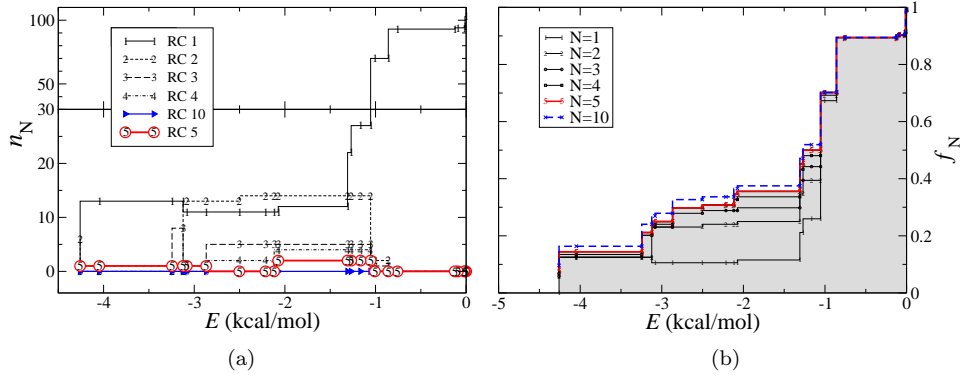


Figure 3. (a) The number n_N of C_α atoms contained within rigid clusters (RC) $N = 1, \dots, 5$ and 10 of the 1HRC structure. Smaller, higher-numbered clusters do not contain more than one C_α . (b) The fraction f of the protein's C_α atoms contained within clusters 1 to N . The line corresponding to the $N = 5$ data has been shaded to show that the inclusion of rigid clusters 1 through 5 captures the large-scale rigidity of the protein.

for, e.g. those C_α lying within rigid clusters $N = 1$ to 5 and also 10. The inclusion of the first five rigid clusters captures the large-scale rigidity of the protein; the difference between $N = 5$ and $N = 10$ is minimal. We therefore use the $N = 5$ measure, $f_5(E)$, to quantify protein rigidity hereafter, which we will refer to as *mainchain rigidity*. We emphasize that we have also computed all results presented here for $N = 4$ and $N = 6$ with quantitatively similar and qualitatively identical results.

It is worth noting the "stepped" appearance of our graphs. This is because a given pattern of rigidity persists as the cutoff is lowered until at a specific value it changes and a certain amount of rigidity is lost.

2.4. Structural comparison by RMSD

When dealing with slightly varying crystal structures of the same protein, we quantify the structural variation by aligning the C_α atoms of two structures and obtaining the root-mean-square deviation between C_α positions,

$$d = \sqrt{\frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} d_{ii}^2} \quad (2)$$

where d_{ii} is the distance between the C_α atoms of residue i in the aligned structures.

3. Results and discussion

3.1. Comparing rigidity of very similar proteins: cytochrome C

In Figure 4 we show dilution plots for four mitochondrial cytochrome C structures obtained from horse crystallised under different conditions as detailed in Table 1. The structural variations between these four structures are small (Table 2a), the largest being 0.572\AA between 1U75 and 1WEJ; for comparison, Minary and Levitt [36] consider structures within $d \simeq 4\text{\AA}$ as "near-native".

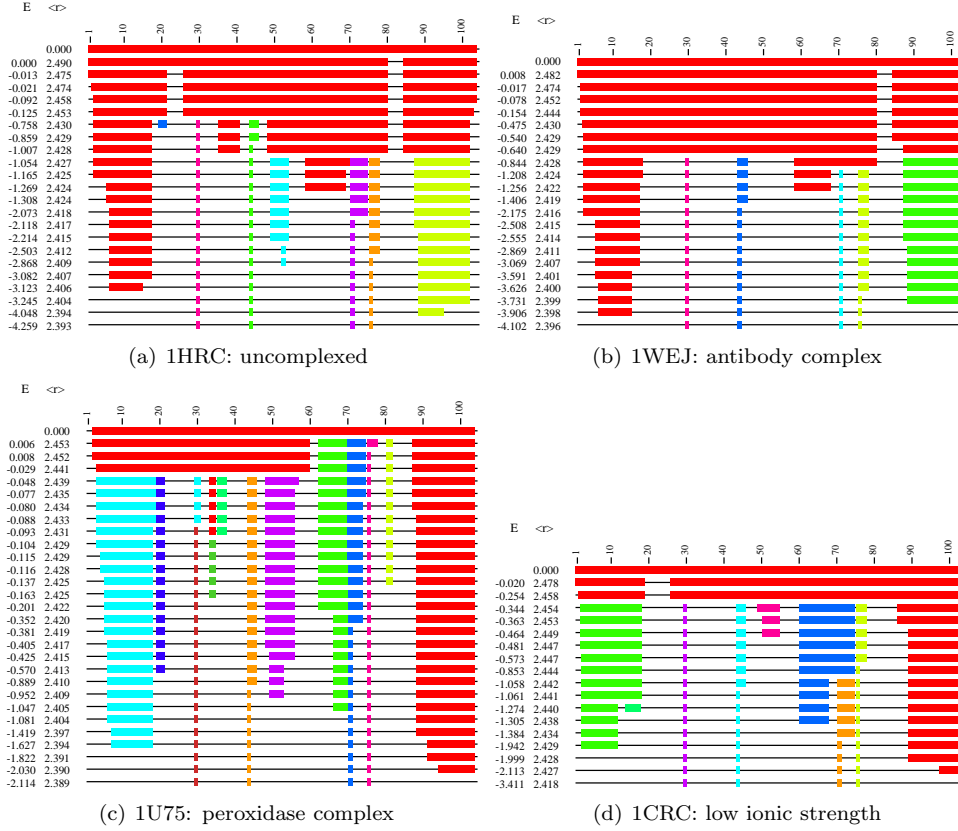


Figure 4. Dilution plots for four crystal structure of horse cytochrome C. The four structures are very similar to each other (see text) and display similar patterns of rigidity loss. The central portion of the protein sequence breaks up into smaller clusters (e.g. close to $E=-1$ for 1HRC and $E=-0.7$ for 1WEJ) and then becomes entirely flexible, while the rigidity of the two ends of the sequence, around residues 5–15 and 90–105, persists longer; these portions are α -helical in secondary structure.

(a)	From\To:	1HRC	1CRC	1WEJ
	1CRC	0.32	—	—
	1WEJ	0.318	0.321	—
	1U75	0.472	0.53	0.572

(b)	From\To:	5CYT	1I55	1I54
	1I55	0.27	—	—
	1I54	0.2668	0.041	—
	1LFM	0.286	0.116	0.087

Table 2. Root-mean-square deviation in Å for C_{α} positions among (a) four horse cytochrome C structures and (b) four tuna cytochrome C structures, showing the similarity of the structures.

The patterns of rigidity loss in Figure 4 appear quite similar on first inspection. The central portion of the protein sequence breaks up into smaller clusters and then becomes entirely flexible, while the rigidity of the two ends of the sequence, around residues 5–15 and 90–100, persists longer; due to this persistence, these portions (α -helical in secondary structure) were identified in [22] as being the folding core of

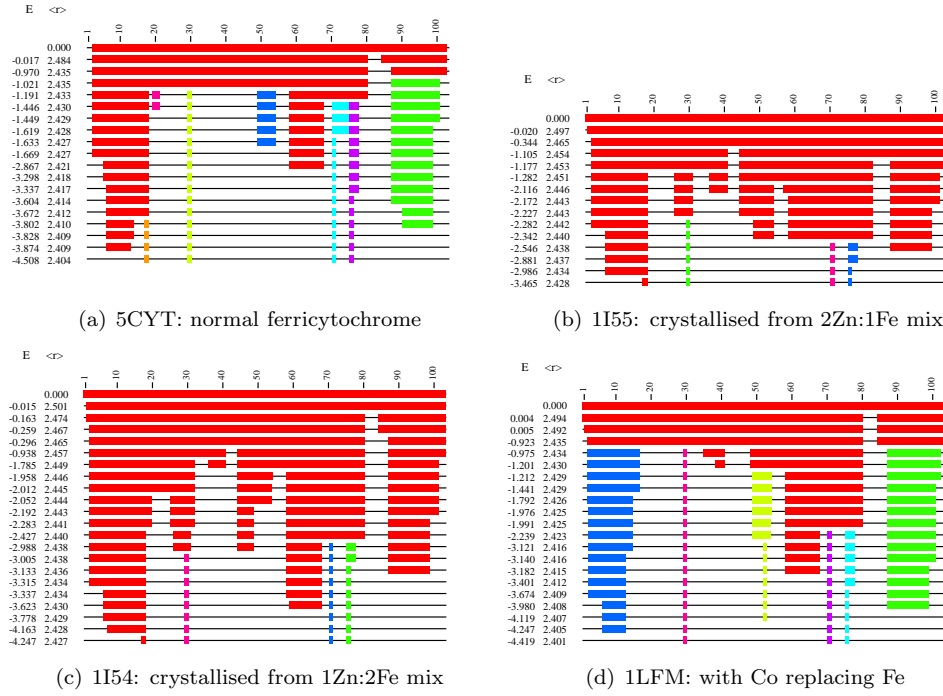


Figure 5. Rigidity dilutions for four forms of tuna cytochrome C crystallised with different metal ion content in the heme groups. (a) normal Fe, (b) from a mixture with 2Zn:1Fe, (c) from a mixture with 2Fe:1Zn, (d) with Co.

cytochrome C, in agreement with experimental evidence.

On closer inspection, however, we can see differences between the four structures in the cutoff energies in which changes in rigidity take place. For example, in structures 1HRC and 1WEJ, the terminal α -helical sequences remain rigid down to cutoff values below -3 kcal/mol, while in 1CRC and 1U75 these sequences are already largely flexible at a cutoff value of -2 kcal/mol. We plot the mainchain rigidity of these four proteins as a function of cutoff energy during dilution in Figure 6(a). The differences in energy scale of the rigidity loss is now clearly visible. Note in particular that in the energy range around -0.1 to -0.6 kcal/mol, two of the structures retain mainchain rigidity ($f_5 > 0.9$) while the other two have already dropped to $f_5 < 0.5$.

In Figure 5, we now consider mitochondrial cytochrome C structures (from tuna) which differ only in their heme-group metal content and are structurally very similar (RMSD values given in Table 2b). We see that the dilution plots for the tuna protein have similar shapes and indeed are quite similar to those for the horse protein (Figure 4). There are differences, however: in particular, in structure 1I54 the α -helical region at residues 60–70 remains rigid to lower cutoff values than that at residues 90–100, which would disagree with the “folding core” prediction of reference [22]. We would therefore argue that physical conclusions drawn from rigidity analysis should be based on the comparison of as many structures as possible if they are to be robust.

Once we plot mainchain rigidity as a function of cutoff energy we again observe differences in the energy scales at which rigidity is lost, (Figure 6b). The greatest

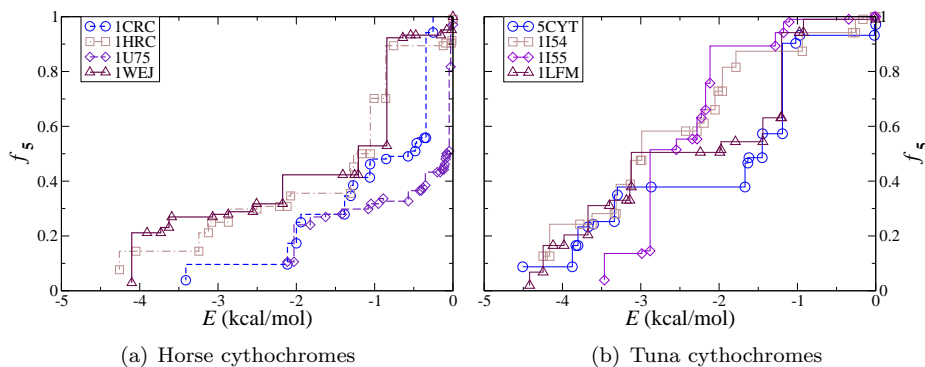


Figure 6. (a) Mainchain rigidity as a function of hydrogen bond energy cutoff E during dilution for four horse mitochondrial cytochrome C structures. Note that for cutoff energy values in the region of -0.5 kcal/mol, structure 1HRC and 1WEJ are almost completely rigid while structures 1U75 and 1CRC are less than 50% rigid. (b) Mainchain rigidity for four tuna cytochrome C structures. Note the considerable differences in behaviour between, for example, 5CYT and 1I55 in the -1 to -2 kcal/mol energy range, even though the structures differ from each other only slightly.

discrepancy appears in the energy range from -1 to -2 kcal/mol; here the 5CYT structure has $f_5 \simeq 0.4$ while 1I55 has $f_5 \simeq 0.9$, although the structures differ by less than $d = 0.3\text{\AA}$ in C_α RMSD.

3.2. Variability of energy scales and selection of cutoff values

It is clear from our investigation of cytochrome C structures that the rigidity analysis at a given cutoff value on very similar structures can easily produce different results. This is not because the dilution plots for these structures differ drastically in their shape, but rather because the cutoff energy at which a major change in rigidity takes place can differ by approximately 1 kcal/mol between very similar structures. This sensitivity of cutoff energy scales to small structural variations is understandable if we consider, for example, the distance dependence for the hydrogen-bond energy function [19]; we show in Figure 1 that a variation in donor-acceptor distance of only 0.1\AA can shift the hydrogen bond energy by around 1 kcal/mol. Thus while the hydrogen bond energy function is successful in distinguishing weaker from stronger bonds, its resolution is limited to approximately 1 kcal/mol.

This implies that exact values of the hydrogen bond cutoff energy cannot easily be transferred between different crystal structures. Rather, it is advisable to perform rigidity dilution on the specific protein structure(s) of interest and to observe how the rigidity changes as the weaker bonds are eliminated, and which portions of the structure are most stable, before selecting an appropriate cutoff value for further investigation of the rigidity/flexibility of the structure(s). While this is an a sense the implicit message of the wide variety of cutoff values used in previous studies (see section Appendix A) we believe the point should be made explicitly for the benefit of potential users of the method.

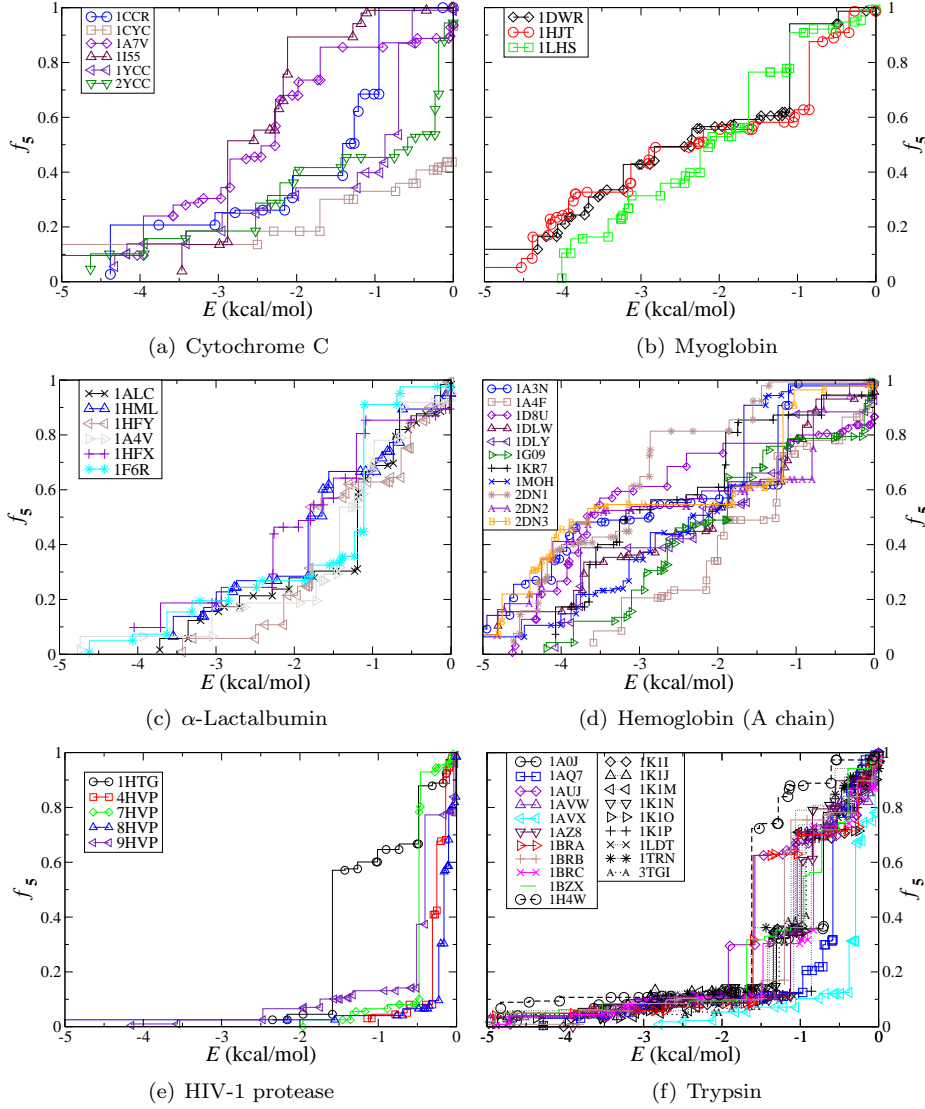


Figure 7. Rigidity dilutions for different families of proteins: cytochrome C, myoglobin, α -lactalbumin, hemoglobin, HIV-1 protease and trypsin. We can see that proteins can display either a “gradual” (a–d) or a “sudden” (e–f) pattern of rigidity loss.

3.3. Patterns of rigidity loss

For the cytochromes that we have so far considered (3.1), the general pattern is one of gradual rigidity loss, particularly for $|E| > 1$. This indicates a hierarchy of stability in the rigid clusters, with some areas being rigidified by very weak hydrogen bonds, some by bonds of medium strength and some by the strongest bonds. This is reminiscent of the gradual or second-order rigidity transition observed in some glassy networks [31], specifically those with a wide diversity in their constraint distribution.

Glassy networks with less diverse constraint networks, however, show a sudden, first-order-like rigidity transition in which the structure passes between largely flexible and largely rigid states on the addition or removal of only a few constraints.

In Figure 7 we show the patterns of rigidity loss for six different families of proteins as listed in Table 1. Our sample falls into two classes, those displaying a gradual pattern of rigidity loss (Figure 7, (a) cytochrome C, (b) myoglobin, (c) lactalbumin, and (d) hemoglobin) and those displaying a sudden loss of rigidity once weak bonds are eliminated (Figure 7, (e) HIV-1 protease and (f) trypsin). For proteins in this second class, all the 25 structures that we examine display rapid loss of mainchain rigidity as weak bonds are removed and the mainchain has become almost entirely flexible once the cutoff energy is reduced below -2 kcal/mol. This indicates that the rigidity of clusters in these proteins is due to weaker hydrogen bonds and we do not see (as we do in the other four proteins) the persistence of rigid clusters bound by stronger hydrogen bonds.

The HIV-1 protease is a natural homodimer and we consider the rigidity of the dimer, as in [19]. For the other protein families our data is obtained from single protein chains; for example, for hemoglobin we analyse α -hemoglobin chains. It should be clear that a protein chain treated in isolation always has fewer constraints than when treated as part of a complex, and indeed we find that the individual chains from the HIV-1 protease structure are even less rigid than the entire dimer (data in Supplementary Materials). For the case of hemoglobin, we can confirm that the rigidity of the isolated A-chain seen in Figure 7(d) differs only slightly from the rigidity of the same chain when analysed as part of the full tetrameric hemoglobin structure (data in Supplementary Materials). Consideration of isolated HIV-1 protease monomers, or of full hemoglobin tetrameric complexes, thus does not alter their classification in terms of gradual or sudden loss of rigidity.

Comparison of these six protein families thus leads us to the conclusion that protein structures, like glassy networks, can display two distinct patterns of rigidity loss depending on the diversity of their constraint networks. We have identified two families of proteins, HIV protease and trypsin, whose members display rapid loss of rigidity as weaker hydrogen bonds are eliminated, in contrast to four other families of proteins which display a gradual loss of rigidity indicating a hierarchy of hydrogen-bond strengths in the constraints that maintain protein rigidity.

4. Conclusion and outlook

Our motivation in this study was twofold: to clarify a methodological issue in the use of rigidity analysis on protein structures, by determining the robustness of RCDs against small structural variations and the significance of the cutoff energy value, and to obtain an insight into the patterns of rigidity loss during hydrogen-bond dilution, by comparison with the observed patterns in glassy networks.

On the first point, we find that there is considerable variation in the RCDs of structurally similar proteins during dilution. Figure 6, for example, shows that among a group of cytochrome C structures drawn from similar eukaryotic mitochondria, energy cutoffs in the range from 0 to -2 kcal/mol (such as have typically been used for FIRST/FRODA simulations of flexible motion [24,25,27,28]) can produce a wide range of degrees of mainchain flexibility. We conclude that the results of rigidity analysis on individual crystal structures should not be over-interpreted as being “the” RCD for a protein. The hydrogen-bond energy function in FIRST is quite sensitive to small

structural variations; while it successfully divides weaker from stronger bonds, it is not possible to identify a unique value of the hydrogen bond cutoff energy which can be applied to all protein structures to give meaningful results. Rather, each protein structure should first be subjected to rigidity dilution to produce a dilution plot; a suitable value of the cutoff energy can then be chosen to test a specific hypothesis about the rigidity and flexibility of the protein. Similarly, when physical significance is attached to the pattern of rigidity loss [22], then multiple similar examples of a given protein structure should be studied in order to be robust against structural variation.

On the second point, we find that proteins can display either gradual (second-order-like) or sudden (first-order-like) patterns of rigidity loss during dilution. We find sudden rigidity loss in two proteases, eukaryotic trypsin and viral HIV-1 protease. Both consist largely of β -sheet secondary structure with little α -helical content compared to the other proteins in our set, which may account for their different rigidity behaviour. Previous work [21] has emphasised the analogy between the rigidity transitions of proteins and of glassy networks; we have now found that the two distinct patterns of rigidity transition recently identified in glassy networks [31] are also seen in proteins.

Our results in this paper suggest several avenues for further enquiry. The rigidity of protein monomers extracted from complexes should be systematically compared with their rigidity within the complex, which will be affected by interchain interactions. The robustness of flexible motion simulations based on rigidity analysis using different cutoff values must also be investigated. A recent study of the flexible motion of myosin [29] found that the flexible motion of the myosin structure appeared qualitatively similar over a wide range of cutoff values covering both highly flexible and more rigid structures. This suggests that rigidity analysis retains its value as a natural coarse-graining for simulations even if the rigidity behaviour during dilution is as variable as we have found.

Acknowledgments

We thankfully acknowledge discussions with R. Freedman and T. Pinheiro. We gratefully acknowledge financial support from the Leverhulme Trust (SAW and RAR, grant F/00 215/AH), BBSRC (JEJ) and EPSRC (JEJ and RAR, EP/C007042/1). We would like to thank two anonymous reviewers for their valuable comments and perspective.

Appendix A. Cutoff values in previous studies using FIRST

Jacobs *et al.* [19] comment that the results of FIRST analysis should not be sensitive to the typical “fluctuations known to occur within protein structures”. Their advice is that the cutoff should be at least -0.1 kcal/mol in order to eliminate a large number of very weak hydrogen bonds with energies in the range from 0.0 to -0.1 kcal/mol, and that a natural choice is near the “room temperature” energy of -0.6 kcal/mol. As we have seen in section 3.2, this criterion is not sufficient to avoid sensitivity to small structural variations.

Rader *et al.* [21] consider the protein folding transition by monitoring $\langle r \rangle$ (mean number of bonded neighbours per atom) during rigidity dilution; they do not, however, comment on the hydrogen bond energy values. Hespenheide *et al.* [22] identify the protein folding core with “the set of secondary structure that remain rigid the

longest in the simulated denaturation”, without regard to the exact values of the cutoff energy at which rigidity is lost. Here the cutoff energy is used qualitatively to distinguish *weaker* from *stronger* bonds. In considering the rigidity of virus capsid protein complexes, Hespenheide *et al.* [20] make use of a cutoff of -0.35 kcal/mol, a value chosen so that capsid protein dimers would be flexible while the inner ring of proteins in a pentamer of dimers would be rigid, and draw conclusions about the rigidity of other multimeric complexes. Meanwhile, Hemberg *et al.* [26] use a different cutoff of -0.7 kcal/mol in a study on the dynamics of capsid assembly.

The FRODA geometric simulation algorithm [24] makes use of the RCD generated by FIRST as a coarse-graining. Simulations of protein mobility using FIRST/FRODA have tended to use cutoff values that are systematically lower than in applications of FIRST alone; typically -1 kcal/mol or lower [24, 25, 27–29], as cutoff values closer to zero seem to include too many constraints to allow large-scale motion to occur. In a paper on the combination of rigidity analysis and elastic network modelling, Gohlke *et al.* [30] discuss RCDs of two protein crystal structures but do not specify a cutoff value, though the FRODA mobility simulations given in Figure 3a of that paper were performed using a cutoff of -1.5 kcal/mol and give an excellent match to experimental data from NMR ensembles.

- [1] Canino LS, Shen T, and McCammon. Changes in flexibility upon binding: application of the self-consistent pair contact probability method to protein-protein interactions. *J. Chem. Phys.*, 117:9927–9933, 2002.
- [2] Case DA. Molecular dynamics and normal mode analysis of biomolecular rigidity. in thorpe m. f., duxbury p. m., eds. rigidity theory and applications. *New York: Kluwer Academic/Plenum Publishers*, pages 329–344, 1999.
- [3] Bahar I, Atilgan AR, and Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.
- [4] Ming D, Kong Y, We Y, , and Ma J. Substructure synthesis method for simulating large molecular complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 100:104–109, 2003.
- [5] Ming D, Kong Y, Wakil SJ, Brink J, and Ma J. Domain movements in human fatty acid synthase by quantized elastic deformational model. *Proc Nat Acad Sci USA*, 99:7895–7899, 2002.
- [6] Tama F, Wriggers W, and Brooks CL 3rd. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Bio.*, 321:297–305, 2002.
- [7] Halle B. Flexibility and packing in proteins. *Proc Natl Acad Sci USA*, 99:1274–1279, 2002.
- [8] Tirion MM. Large amplitude elastic motions in proteins from single-parameter atomic analysis. *prl*, 77:1905–1908, 1996.
- [9] Tama F, Gadea FX, Marques O, and Sanejouand YH. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41:1–7, 2000.
- [10] Delarue M and Sanejouand YH. Simplified normal mode analysis of conformational transitions in dna-dependant polymerases: the elastic network model. *J. Mol. Biol.*, 320:1011–1024, 2002.
- [11] Petrone P and Pande VS. Can conformational change be described by only a few normal modes? *Biophys J.*, 90:p1583–1593, 2006.
- [12] Vihinen M, Torkkila E, and Riikonen P. Accuracy of protein flexibility predictions. *Proteins*, 19:141–149, 1994.
- [13] Holm L and Sander C. Parser for protein folding units. *Proteins*, 19:256–268, 1994.
- [14] Zehfus MH and Rose GD. Compact units in proteins. *Biochemistry*, 25:5759–5765, 1986.
- [15] Karplus PA and Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213, 1985.
- [16] Mayorov V and Abagyan R. A new method for modeling large-scale rearrangements of protein domains. *Proteins*, 27:410–424, 1997.
- [17] Flores M, Echols N, Milburn D, Hespenheide BM, Keating K, Lu J, Wells SA, Yu EZ, Thorpe MF, and Gerstein M. The database of macromolecular motions: new features added at the decade mark. *Nucleic Acid Research*, 34:D296–D301, 2005.
- [18] Jacobs DJ and Thorpe MF. Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.*, 75:4051–4054, 1995.
- [19] Jacobs DJ, Rader AJ, Kuhn LA, and Thorpe MF. Protein flexibility predictions using graph

- theory. *PROTEINS: Struct., Func. and Gen.*, 44:150–165, 2001.
- [20] Hespenheide BM, Jacobs DJ, and Thorpe MF. Structural rigidity and the capsid assembly of cowpea chlorotic mottle virus. *J. Phys.: Condens. Matter*, 16:S5055–S5064, 2004.
 - [21] Rader AJ, Hespenheide BM, Kuhn LA, and Thorpe MF. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci.*, 99:3540–3545, 2002.
 - [22] Hespenheide BM, Rader AJ, Thorpe MF, and Kuhn LA. Identifying protein folding cores: Observing the evolution of rigid and flexible regions during unfolding. *J. Mol. Graph. & Model.*, 21:195–207, 2002.
 - [23] Thorpe MF, Lei M, Rader AJ, Jacobs DJ, and Kuhn LA. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. & Model.*, 19:60–69, 2001.
 - [24] Wells SA, Menor S, Hespenheide BM, and Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.*, 2:S127–S136, 2005.
 - [25] Jolley CC, Wells SA, Hespenheide BM, Thorpe MF, and Fromme P. Docking of photosystem i subunit c using a constrained geometric simulation. *J. Am. Chem. Soc.*, 128:8803–8812, 2006.
 - [26] Hemberg M, Yaliraki SN, and Barahona M. Stochastic kinetics of viral capsid assembly based on detailed protein structures. *Biophysical journal*, 90:3029–3042, 2006.
 - [27] Jolley CC, Wells SA, Fromme P, and Thorpe MF. Fitting low-resolution cryo-em maps of proteins using constrained geometric simulations. *Biophys. J.*, 94:1613–1621, 2008.
 - [28] Macchiarulo A, Nuti R, Bellochi D, Camaioni E, and Pellicciari R. Molecular docking and spatial coarse graining simulations as tools to investigate substrate recognition, enhancer binding and conformational transitions in indoleamine-2,3-dioxygenase (ido). *Biochim. et Biophys. Acta-Proteins and Proteomics*, 1774:1058–1068, 2007.
 - [29] Sun M, Rose MB, Ananthanarayanan SK, Jacobs DJ, and Yengo CM. Characterisation of the pre-force-generation state in the actomyosin cross-bridge cycle. *Proc. Nat. Acad. Sci.*, 105:8631–8636, 2008.
 - [30] Gohlke H and Thorpe MF. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.*, 91:2115–2120, 2006.
 - [31] Sartbaeva A, Wells SA, Huerta A, and Thorpe MF. Local structural variability and the intermediate phase window in network glasses. *Phys. Rev. B*, 75:224204, 2007.
 - [32] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000. <http://www.rcsb.org>.
 - [33] DeLano W. The pymol molecular graphics system. www.pymol.org.
 - [34] Word JM, Lovell SC, Richardson JS, and Richardson DC. Asparagine and glutamine: Using hydrogen atoms contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285:1735–1747, 1999. <http://kinemage.biochem.duke.edu/software/reduce.php>.
 - [35] Dahiyat BI, Gordon DB, and Mayo SL. Automated design of the surface positions of protein helices. *Protein Sci.*, 6:1333–1337, 1997.
 - [36] Minary P and Levitt M. Probing protein fold space with a simplified model. *J. Mol. Biol.*, 375:920–933, 2008.